Feriel Abdi
M2 – BI
Université Paris Cité

Rapport – Projet Court

Programmation 3 et projet tuteuré

Link to the Github Project : https://github.com/ferielab21/Projet_court_Feriel_Abdi

Introduction

The solvent-accessible surface area of amino acids in proteins is a crucial parameter in the study of polypeptide chain folding and stability calculation. In this project, we focused on implementing the Skrake & Rupley algorithm, which calculates the Solvent Accessible Surface (SAS), an essential descriptor in protein studies, particularly in predicting their structure. Their work was motivated by the need for a simple and objective description of protein molecular structures, with an emphasis on the molecular surface and the environment of reactive groups, features closely related to chemical properties. In this report, we extend this method by specifically examining the nature of interactions between atoms. We will present results for the Dithiol alpha melanotropin peptide (1b0q), which is a growth hormone composed of 12 residues.

Materials and methods

Our objective is to calculate the solvent-accessible surface, defined as the set of positions of a sphere modeling a water molecule (with a radius of 1.4 Å) as it traverses the entire surface of the protein.

The concept of solvent-accessible surface comes into play when we consider a probe in contact with the surface of a molecule. The atoms of the solute molecule, especially the atoms of the protein, are represented as spheres, and our goal is to determine whether these spheres can accommodate the probe or not. To model these spheres, each atom of the protein is represented by a sphere composed of N points, uniformly spaced at a distance equal to the Van der Waals radius of the atom.

Materials

Software and Environment:

- Python: The project was developed using Python programming language.
- Operating System: Linux was used as the operating system for this project.
- Jupyter Notebook: Jupyter Notebook was employed as the integrated development environment for writing and running the initial Python scripts.

Python Libraries :

- Pandas
- Biopython
- NumPy
- Matplotlib
- Argparse

Data Source :

The primary source of structural data was the Protein Data Bank (PDB) from which we extracted the file :

- "1b0q.pdb" : Dithiol alpha melanotropin peptide.

Methods

The first step was data extraction. We initiated by extracting atomic coordinates from the PDB file using the `coords_atoms()` function. Then we used the function `find_accessible_points` to calculate the surface accessibility of the atoms. This crucial step involved iterating through each line of the coordinate dataframe. During each iteration: We employed the `find_neighbors` function to identify neighboring atoms around the current atom. The condition for inclusion was that the distance between the atoms must be less than the sum of the current atom's radius and the radius of a water molecule. Subsequently, we utilized the `create_sphere()` function to generate a sphere with 'n' points around the atom of interest. We then computed the distance between each point on the sphere and the center of each neighboring atom. If this distance was found to be less than the sum of the neighboring atom's radius and the radius of a water molecule, we classified the point as non-exposed to the solvent. Conversely, if the distance exceeded this threshold, we considered the point as solvent-exposed.

To determine the accessible surface area of the atom, we calculated the ratio of exposed points to the total points on the sphere. Then used the following formula :

$$Surface = ratio \times 4 \times \pi \times (VDW\ radius)2$$

With the `accessibility_surf()' function, we aggregated the computed surface areas for each atom to calculate the accessible surface area for each residue. Further analysis involved calculating the percentage of accessibility for each residue. This was done using the formula:

$$Percentage = \frac{Accessible\ surface\ of\ residue}{Total\ surface\ of\ residue} \times 100$$

These results are presented in Table 1 within the Results section.

To determine the overall solvent accessibility of the protein, we summed the accessible surfaces of individual residues.

To validate our approach, we compared it with the "rolling ball" Shrake and Rupley method from Biopython. This established method employs a sphere with a radius equivalent to that of a solvent molecule to probe the surface of the protein.

The detailed findings of this comparative analysis are presented in the Results section, shedding light on the effectiveness and consistency of our methodology relative to the widely accepted Shrake and Rupley approach.

Results

Table 1 : Accessible surfaces and percentages of accessibility of each residue using 100 sphere points

Residue_nb	Residue	Accessible_surface (A^2)	Percentage_accessibility (%)
1	CYS	106.08781400054299	13.728828718949465
2	GLU	148.5659165882613	13.175638025186673
3	HIS	246.6941914416692	19.165576491262325
5	ARG	181.56771909569136	13.03563695416817
6	TRP	228.6413434170808	11.984389408510078
7	CYS	106.04320338486203	13.723055657193969
8	LYS	128.7424669441097	10.89661774090619
9	PRO	127.88544046821042	13.743146522619853
10	VAL	193.96570034381813	20.84442943956786

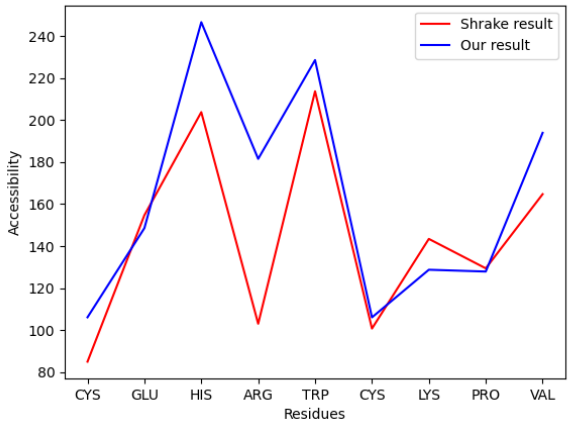

Figure 1: Comparison of the residues' accessible surfaces between our method and biopython

The total accessibility surface of the protein is : **1468.19 A^2** for our method, and **1572.55 A^2** for the biopython's, with a percentage of identity of **83.49%**. What we observe in *Table 1* and *Figure 1* is that our method, based on Shrake and Rupley, closely aligns with the Biopython approach, yielding similar results in terms of accessible surface areas for closely positioned residues. Specifically, for residues such as GLU, TRP, CYS, and PRO, our results closely match. However, we note slightly higher results for residues ARG and HIS. Nevertheless, overall, our findings demonstrate consistency and similarity to the existing method.

Conclusion and perspectives

In this project, we delved into the calculation of solvent-accessible surface areas (SAS) for amino acids within proteins, a crucial parameter in understanding protein folding and stability. We implemented the Skrake & Rupley algorithm, which offers a simple and objective way to describe molecular structures, emphasizing surface characteristics and the environment of reactive groups. Our analysis centered on the Dithiol alpha melanotropin peptide (1b0q), a growth hormone composed of 12 residues. We aimed to calculate the solvent-accessible surface by modeling a water molecule's positions as it traverses the entire protein surface. This method, based on the representation of atoms as spheres and employing a sphere packing approach, enabled us to assess the accessibility of each atom.

We compared our approach to the well-established "rolling ball" Shrake and Rupley method from Biopython. The results revealed a remarkable alignment between the two methods, particularly in terms of accessible surface areas for closely positioned residues such as GLU, TRP, CYS, and PRO. While slight disparities were observed for residues ARG and HIS, our findings underscored the overall consistency and similarity between our method and the existing approach.

In summary, our implementation of the Skrake & Rupley algorithm provides a robust and effective means to calculate solvent-accessible surface areas, contributing to the broader understanding of protein structures and their functional properties. To go futher, we could calculate the relative accessibility of the protein and optimize the code to be able to adapt to bigger proteins.

Bibliography

Shrake & Rupley. Environment and exposure to solvent of protein atoms, lysozyme and insulin. Journal of Molecular Biology, 1973.